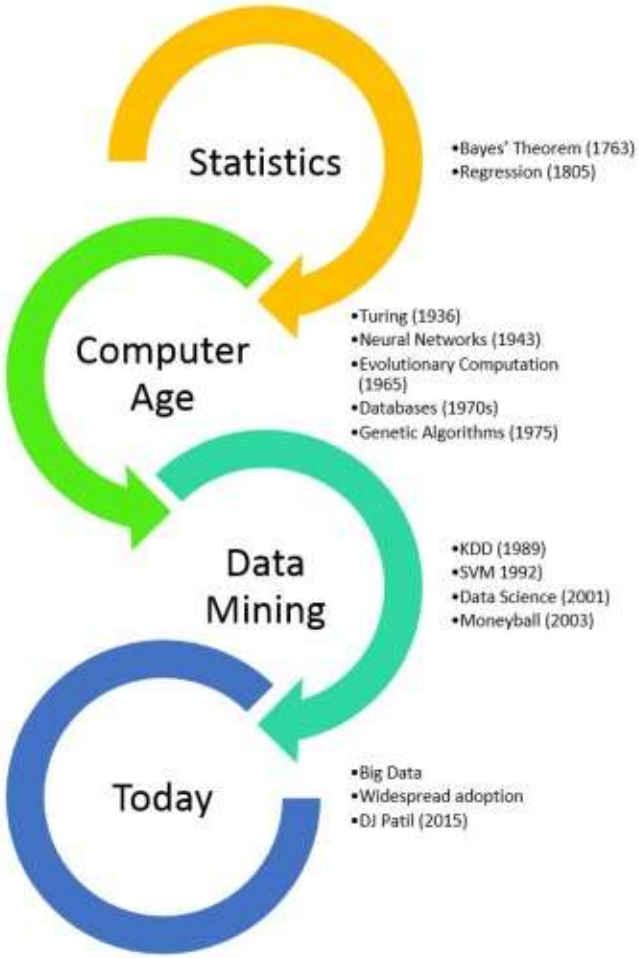
Explanatory Data Analysis

Dr. Tran Anh Tuan,

Faculty of Mathematics and Computer Science,

University of Science, HCMC



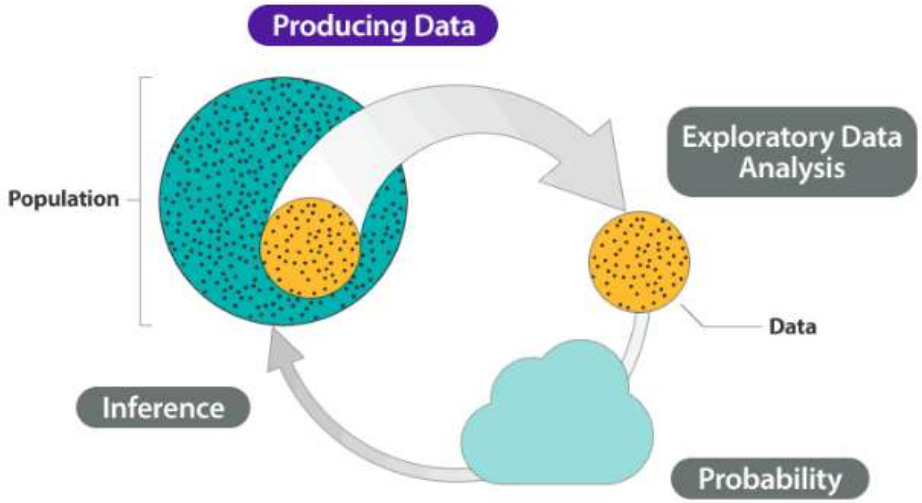
Evolutionary Step	Business Question	Enabling Technologies	Product Providers	Characteristics
Data Collection (1960s)	"What was my total revenue in the last five years?"	Computers, tapes, disks	IBM, CDC	Retrospective, static data delivery
Data Access (1980s)	"What were unit sales in New England last March?"	Relational databases (RDBMS), Structured Query Language (SQL), ODBC	Oracle, Sybase, Informix, IBM, Microsoft	Retrospective, dynamic data delivery at record level
Data Warehousing & Decision Support (1990s)	"What were unit sales in New England last March? Drill down to Boston."	On-line analytic processing (OLAP), multidimensional databases, data warehouses	Pilot, Comshare, Arbor, Cognos, Microstrategy	Retrospective, dynamic data delivery at multiple levels
Data Mining (Emerging Today)	"What's likely to happen to Boston unit sales next month? Why?"	Advanced algorithms, multiprocessor computers, massive databases	Pilot, Lockheed, IBM, SGI, numerous startups (nascent industry)	Prospective, proactive information delivery

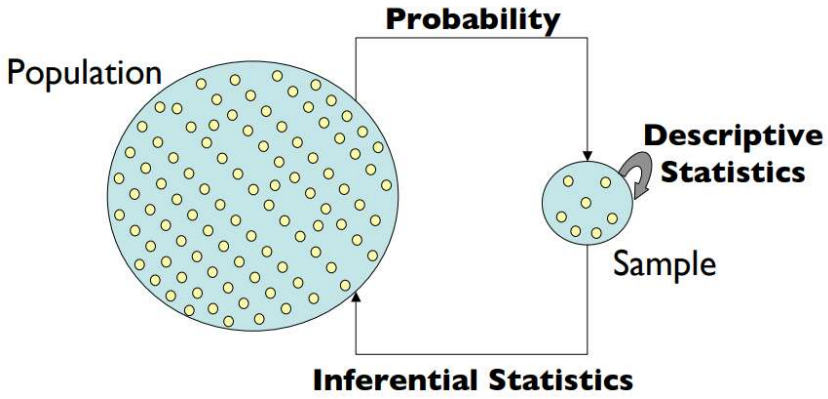
Syllabus

- Lecture 1 : Data Preprocessing
- **Lecture 2 : Explanatory Data Analysis**
- Lecture 3 : Feature Engineering (Feature Importance and Selection)
- Lecture 4 : Association Rule Learning
- Lecture 5 : Unsupervised Clustering
- Lecture 6 : Unsupervised Clustering (cont.)
- Lecture 7 : Anomaly and Outlier Detection
- Lecture 8 : Regression and Classification Learning
- Lecture 9 : Recommendation Learning
- Lecture 10 : Final Project Requirement

What is Exploratory Data Analysis (EDA)

- Exploratory data analysis (EDA) is a crucial component of data science which allows you to develop the gist of what your data look like and what kinds of questions might be answered by them.
- Ultimately, EDA is important because it allows the investigator to make critical decisions about what is interesting to pursue and what probably isn't worth following up on and thus building a hypothesis using the relationships between variables.



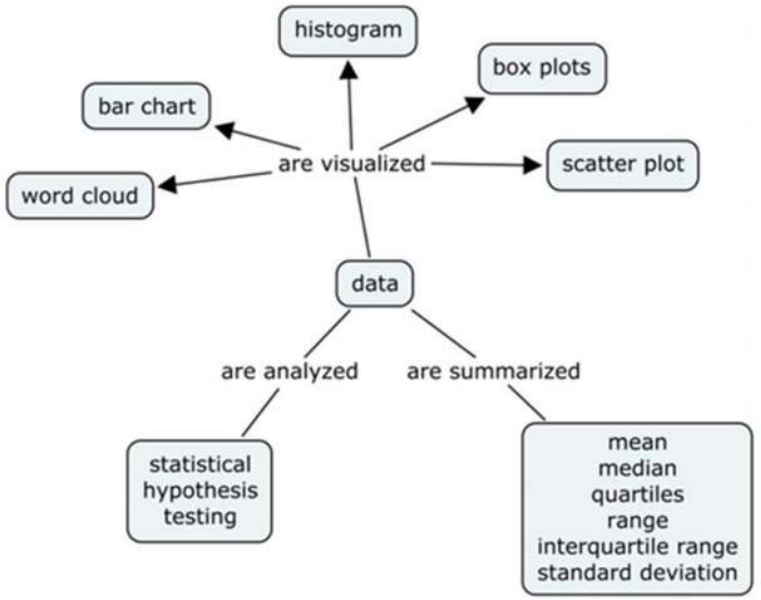


What is Exploratory Data Analysis (EDA)

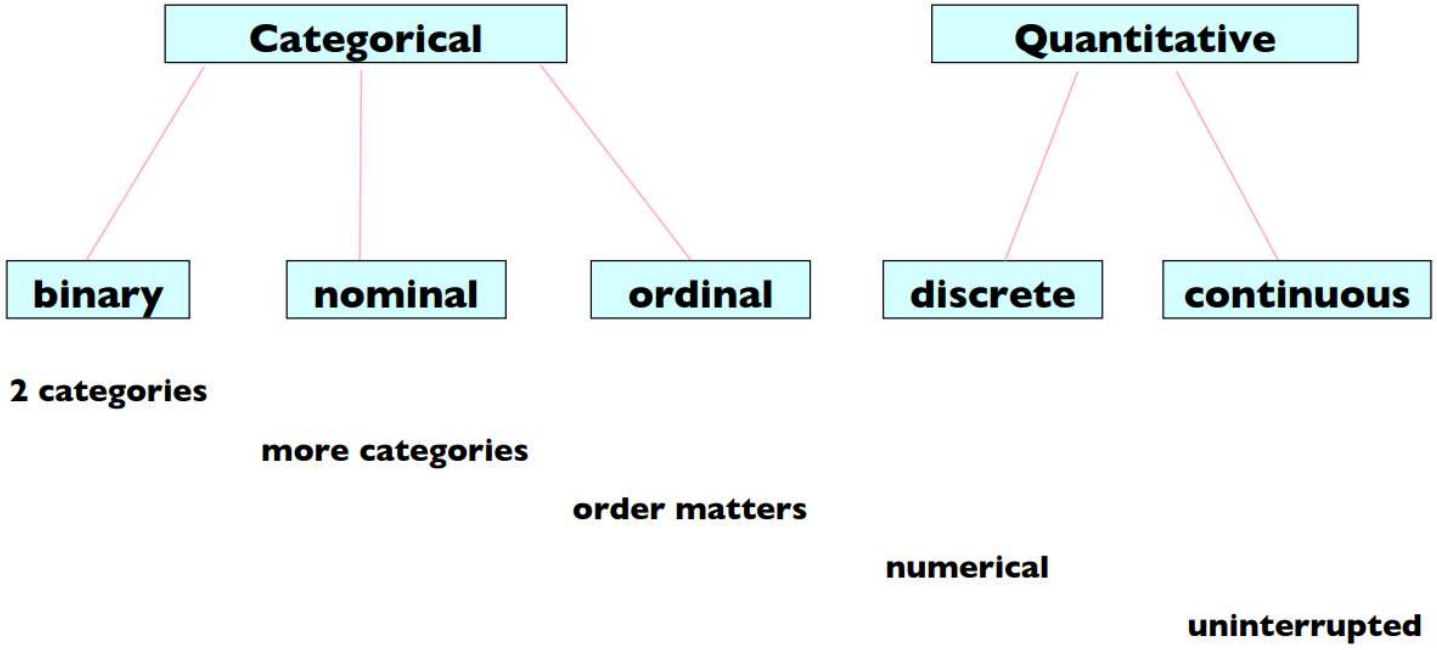
- 1. Exploratory data analysis or “EDA” is a critical first step in analyzing the data from an experiment.
- 2. A statistical model can be used or not, but primarily EDA is for seeing what the **data** can tell us beyond the formal modeling or hypothesis testing task.
- 3. The four types of EDA are univariate non-graphical, multivariate nongraphical, univariate graphical, and multivariate graphical.


Understand the Data
Clean The Data
Analysis of Relationship between variables
Steps Involved In Exploratory Data Analysis
It follows a systematic set of steps to explore the data in the most efficient way possible

Concept Map


histogram
box plots
bar chart
are visualized
scatter plot
word cloud
data
are analyzed
are summarized
statistical hypothesis testing
mean
median
quartiles
range
interquartile range
standard deviation

Types of Data



Dimensionality of Data Sets

- **Univariate:** Measurement made on one variable per subject
- **Bivariate:** Measurement made on two variables per subject
- **Multivariate:** Measurement made on many variables per subject

Numerical Summaries of Data

- ***Central Tendency measures***. They are computed to give a "center" around which the measurements in the data are distributed.

- ***Variation or Variability measures***. They describe "data spread" or how far away the measurements are from the center.

- ***Relative Standing measures***. They describe the relative position of specific measurements in the data.

Descriptive Statistics

- A ***descriptive statistic*** is a summary statistic that quantitatively describes or summarizes features of a collection of information. Descriptive statistics are just descriptive. They do not involve generalizing beyond the data at hand.

For samples:

$$\text{variance} = s^2 = \frac{\Sigma (x - \bar{x})^2}{n-1}$$

$$\text{standard deviation} = s = \sqrt{s^2}$$

Calculating Formula

$$s^2 = \frac{\Sigma x^2 - \frac{(\Sigma x)^2}{n}}{n-1}$$

For populations:

$$\text{variance} = \sigma^2 = \frac{\Sigma (x - \bar{x})^2}{n}$$

$$\text{standard deviation} = \sigma = \sqrt{\sigma^2}$$

Calculating Formula

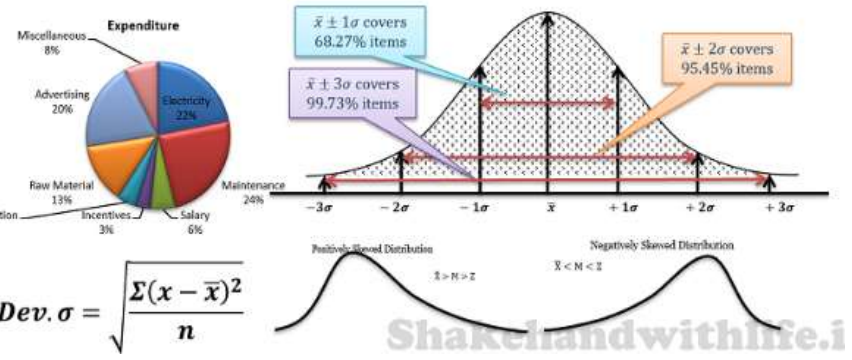
$$\sigma^2 = \frac{\Sigma x^2 - \frac{(\Sigma x)^2}{n}}{n}$$

Mean

Median

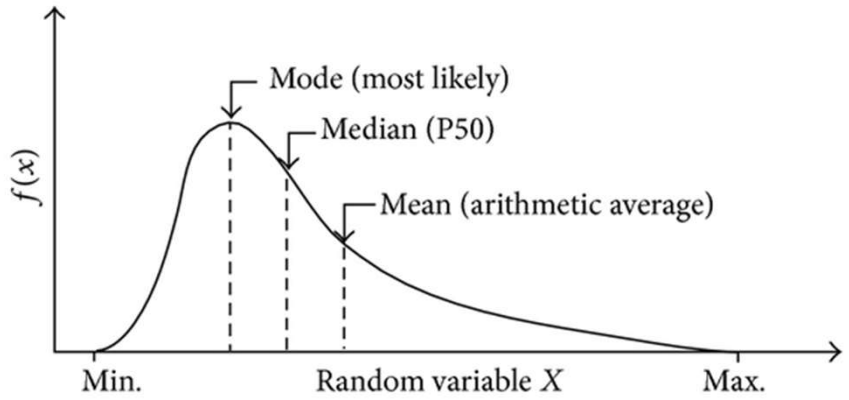
Mode

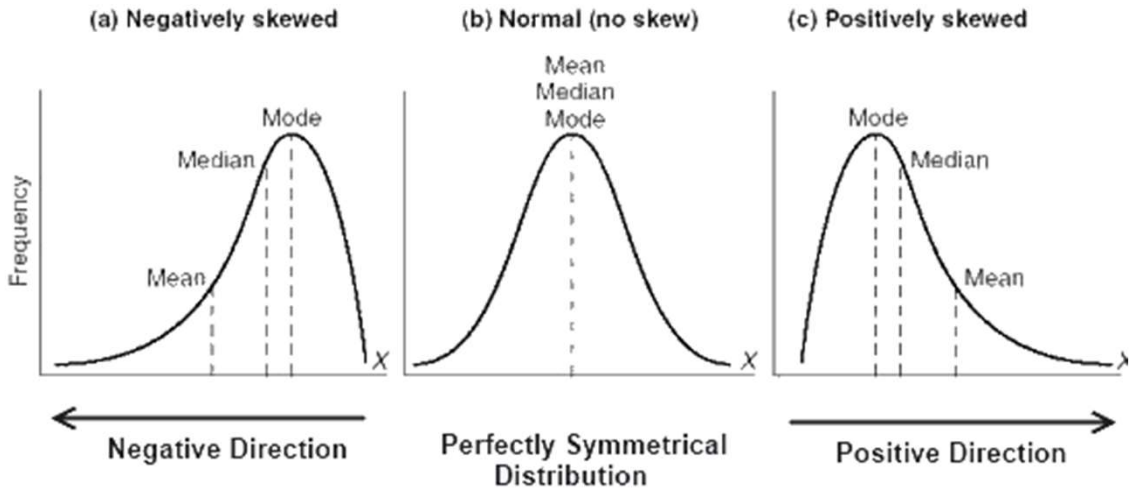
$$\text{Std. Dev. } \sigma = \sqrt{\frac{\Sigma (x - \overline{x})^2}{n}}$$



Descriptive Statistics

- Central tendency
  - Mean, $\overline{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$
  - Median, P_{50} or Q_2
  - Mode, the most frequent value
- Variability (i.e. scale)
  - Range, max minus min
  - Standard deviation, $s = \sqrt{\frac{\sum_{i=1}^{N}(x_i - \overline{x})^2}{N-1}}$
  - Interquartile range, $IQR = Q_3 - Q_1$



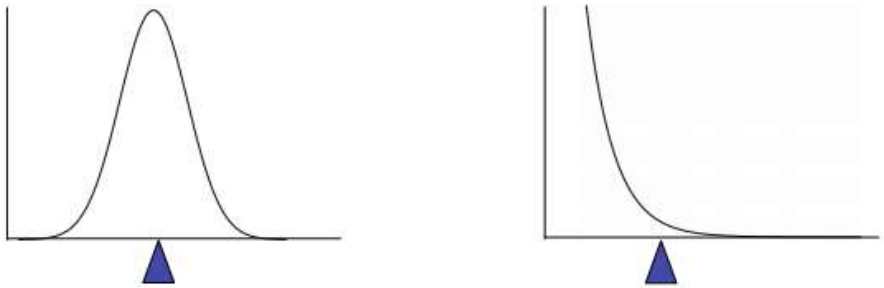


Descriptive Statistics

- Mean

To calculate the average $\overline{x}$ of a set of observations, add their value and divide by the number of observations:

$$\overline{x} = \frac{x_1 + x_2 + x_3 + \ldots + x_n}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i$$



Weighted means:

$$\overline{x} = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i}$$

Trimmed:

$$\overline{x} = \alpha$$

Geometric:

$$\overline{x} = \left(\prod_{i=1}^{n} x_i\right)^{\frac{1}{n}}$$

Harmonic:

$$\overline{x} = \frac{n}{\sum_{i=1}^{n} \frac{1}{x_i}}$$

Descriptive Statistics

- **MEDIAN**
- Median is the value which occupies the middle position when all the observations are arranged in an ascending/descending order. It divides the frequency distribution exactly into two halves. Fifty percent of observations in a distribution have scores at or below the median. Hence median is the 50th percentile. Median is also known as 'positional average'.
- ***Advantages***
  - It is easy to compute and comprehend.
  - It is not distorted by outliers/skewed data.
  - It can be determined for ratio, interval, and ordinal scale.

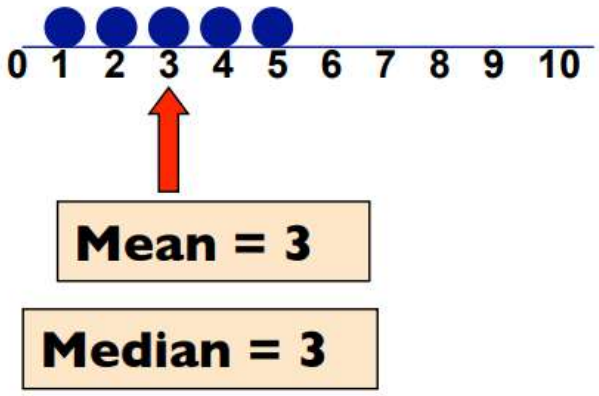
- ***Disadvantages***
  - It does not take into account the precise value of each observation and hence does not use all information available in the data.
  - Unlike mean, median is not amenable to further mathematical calculation and hence is not used in many statistical tests.
  - If we pool the observations of two groups, median of the pooled group cannot be expressed in terms of the individual medians of the pooled groups.

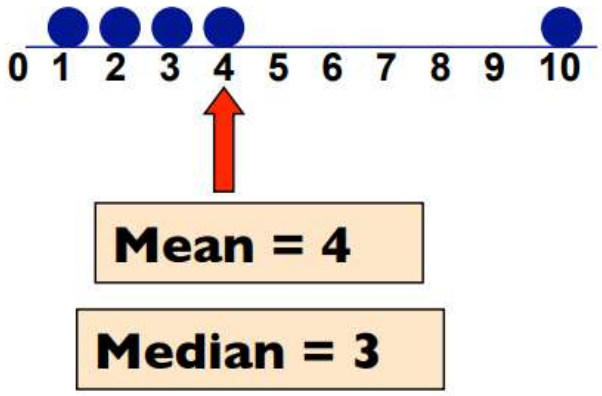
Descriptive Statistics

- **MODE**
- Mode is defined as the value that occurs most frequently in the data. Some data sets do not have a mode because each value occurs only once. On the other hand, some data sets can have more than one mode. This happens when the data set has two or more values of equal frequency which is greater than that of any other value. Mode is rarely used as a summary statistic except to describe a bimodal distribution. In a bimodal distribution, the taller peak is called the major mode and the shorter one is the minor mode.
- ***Advantages***
- It is the only measure of central tendency that can be used for data measured in a nominal scale.
- It can be calculated easily.
- ***Disadvantages***
- It is not used in statistical analysis as it is not algebraically defined and the fluctuation in the frequency of observation is more when the sample size is small.

Which Location Measure Is Best?

- Mean is best for symmetric distributions without outliers
- Median is useful for skewed distributions or data with outliers

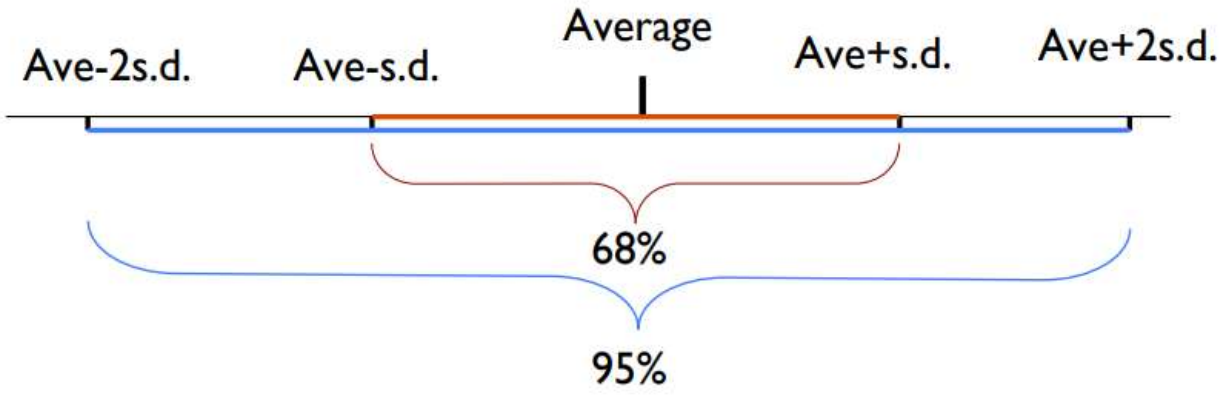




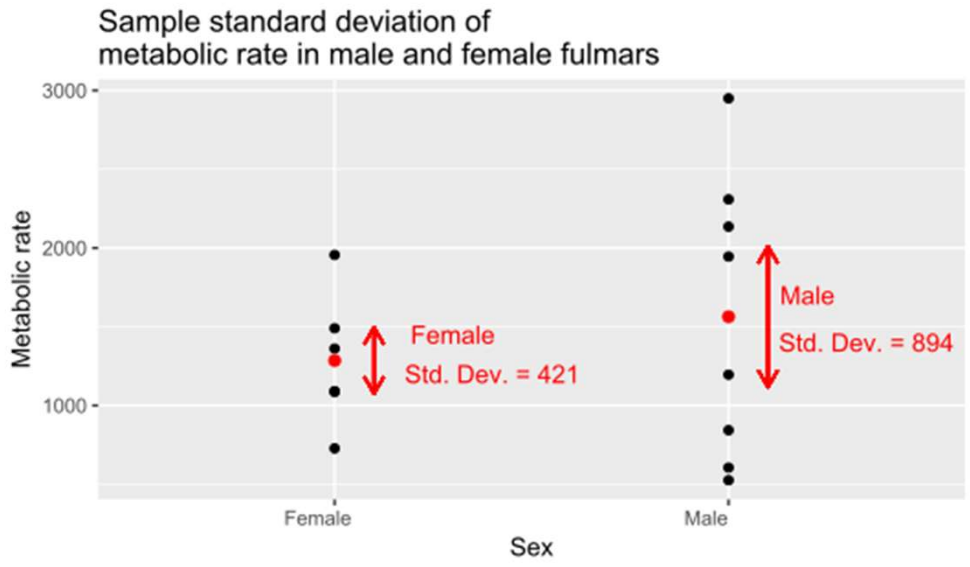
For many lists of observations – especially if their histogram is bell-shaped

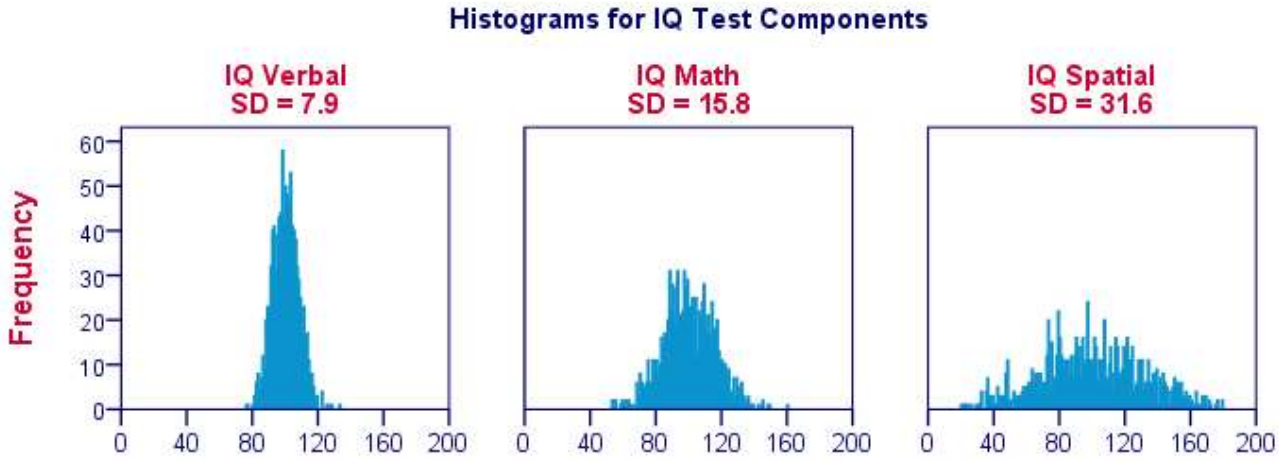
1. Roughly 68% of the observations in the list lie within 1 standard deviation of the average

2. 95% of the observations lie within 2 standard deviations of the average

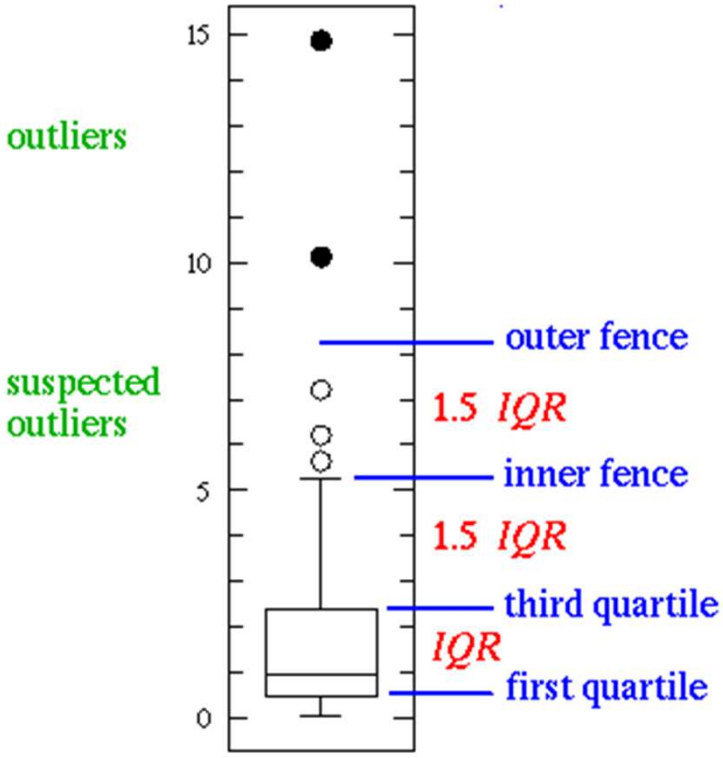


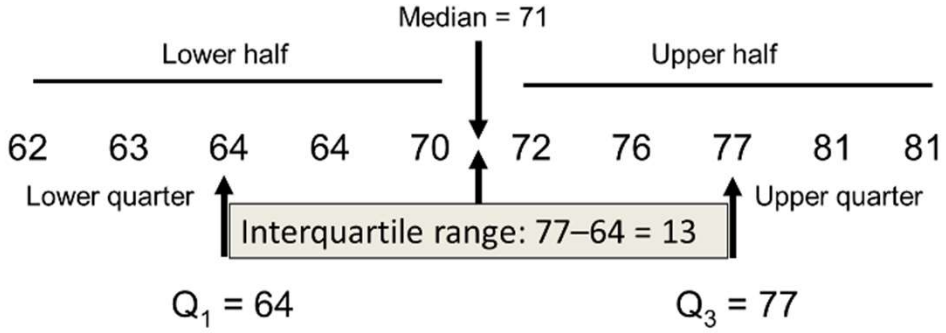
Descriptive Statistics

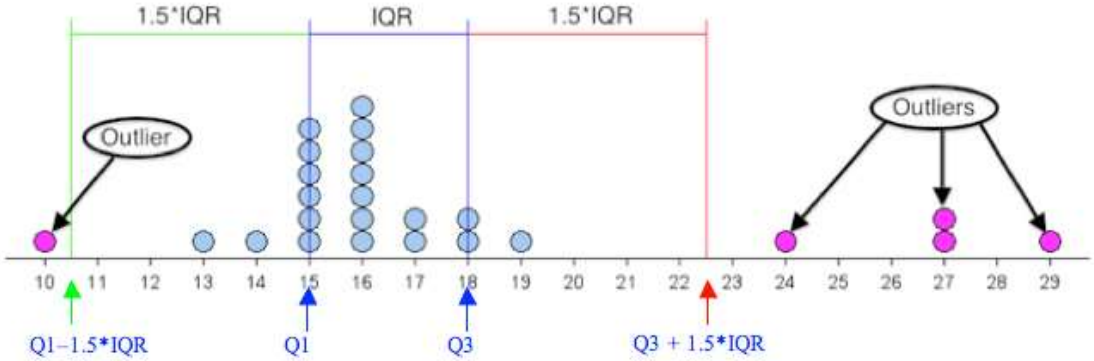




Descriptive Statistics







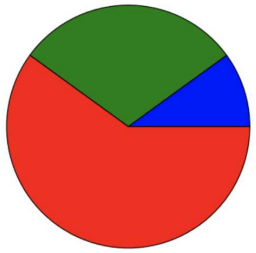
Nominal Data

- Frequencies
  - Count the number of events of interest, f_i
- Proportion (relative frequency)
  - Divide frequency by total number of events, $\frac{f_i}{N}$
- Percentage
  - Multiply proportion by 100, $\frac{f_i}{N} \times 100$
- Illustrate with bar chart or pie chart

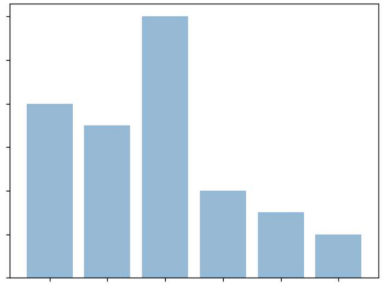
Ordinal Data

- Summarize
  - Frequencies, proportions, and percentages
  - Percentiles ($P_1, P_2, \ldots, P_{99}$)
  - Mode
  - Median
  - Interquartile range
- Illustrate
  - Bar chart
  - Pie chart

Pie Chart

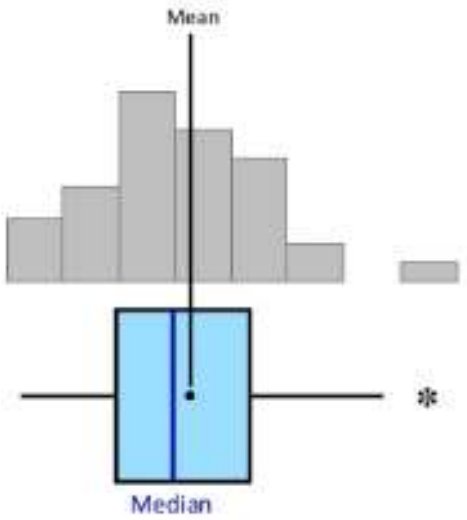


Bar Chart



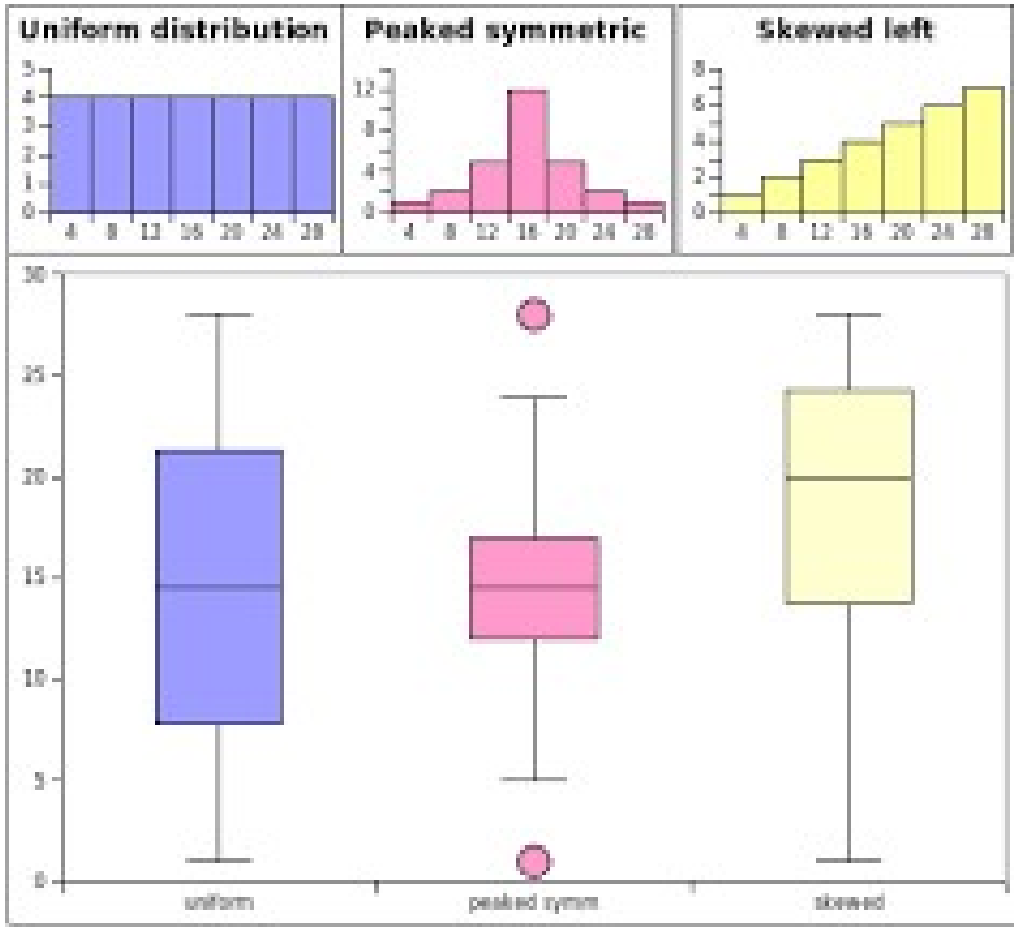
Continuous Data

- Summarize
  - Percentiles, median, interquartile range
  - Mean, median, or mode
  - Standard deviation, range, or IQR
- Illustrate
  - Histogram
  - Boxplot



Continuous Data

Continuous Data

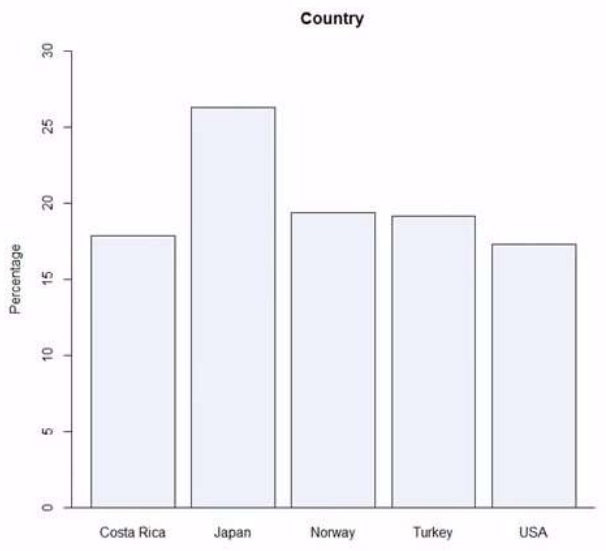


Categorical Summary

Table 1

Country of examinee

Country	Frequency	Percentage
Costa Rica	4,314	17.87
Japan	6,351	26.30
Norway	4,684	19.40
Turkey	4,618	19.13
USA	4,177	17.30

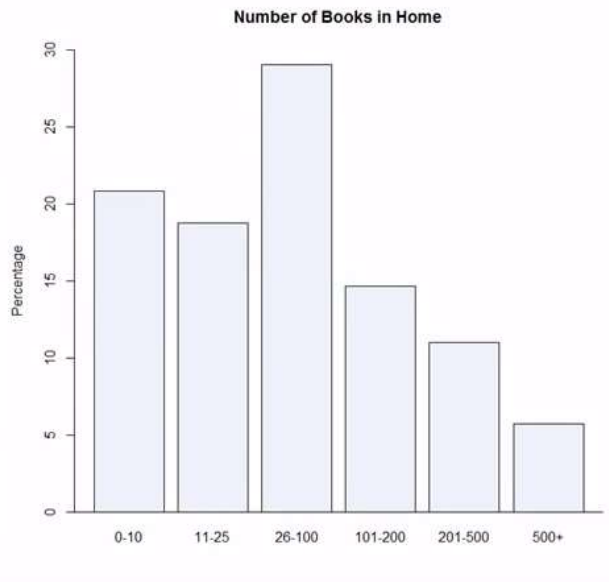


Categorical Summary

Table 2

Number of books in home

Books	Frequency	Percentage
0-10	4,901	20.85
11-25	4,411	18.77
26-100	6,828	29.05
101-200	3,440	14.64
201-500	2,580	10.98
500+	1,344	5.72

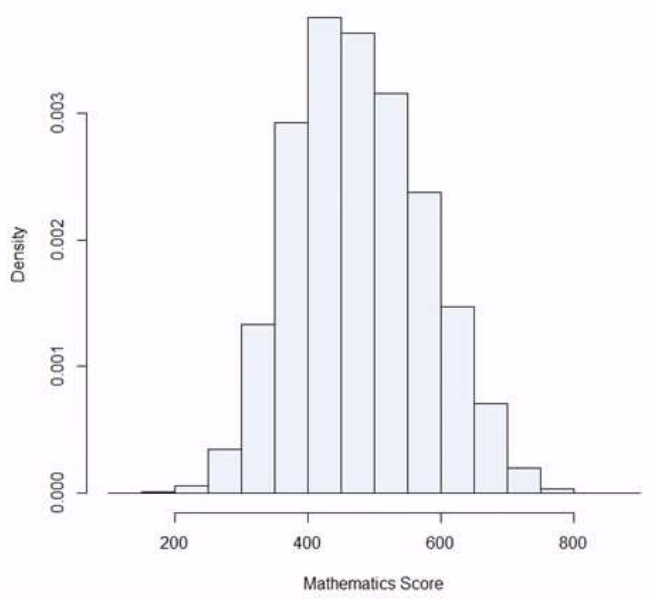


Continuous Summary

Table 3

Mathematics score

Statistic	Value
Min	123.00
Max	851.07
Mean	477.46
S.D.	97.87

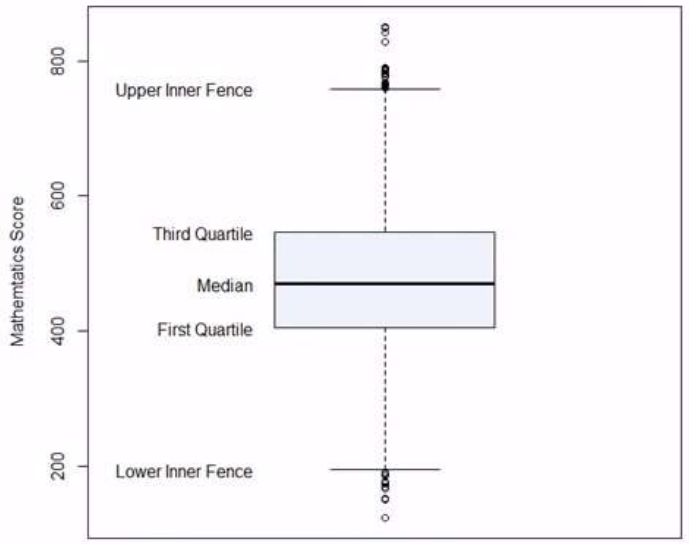


Continuous Summary

Table 3

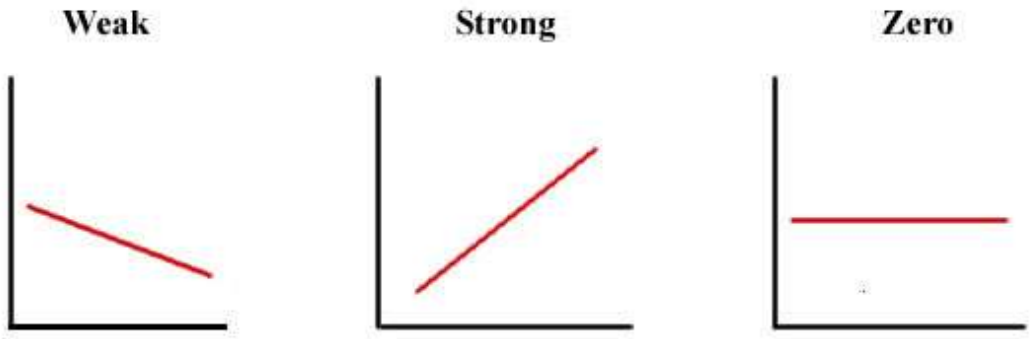
Mathematics score

Statistic	Value
Min	123.00
Max	851.07
Mean	477.46
S.D.	97.87



Relationship between two variables

- How do values of one variable change as values on another variable change?
- Do low scores on one variable correspond to low values on another variable?
- Do large values of one variable correspond to large values on another variable?



Relationship

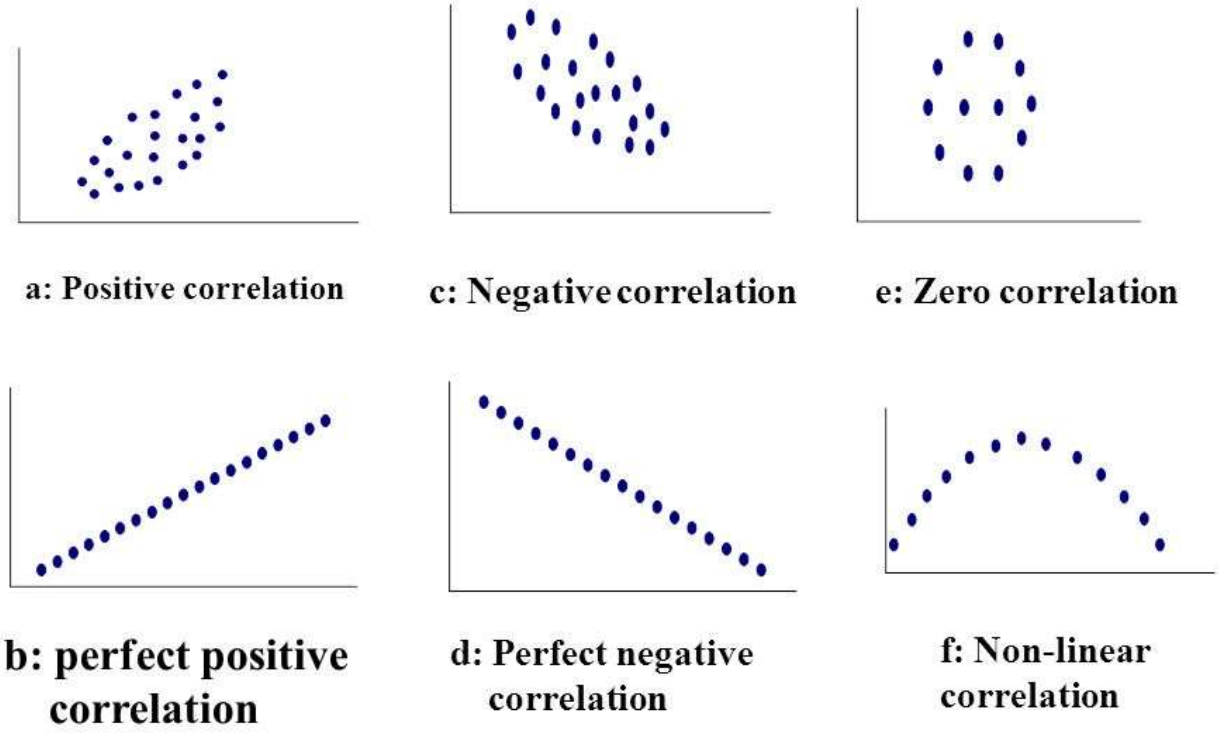


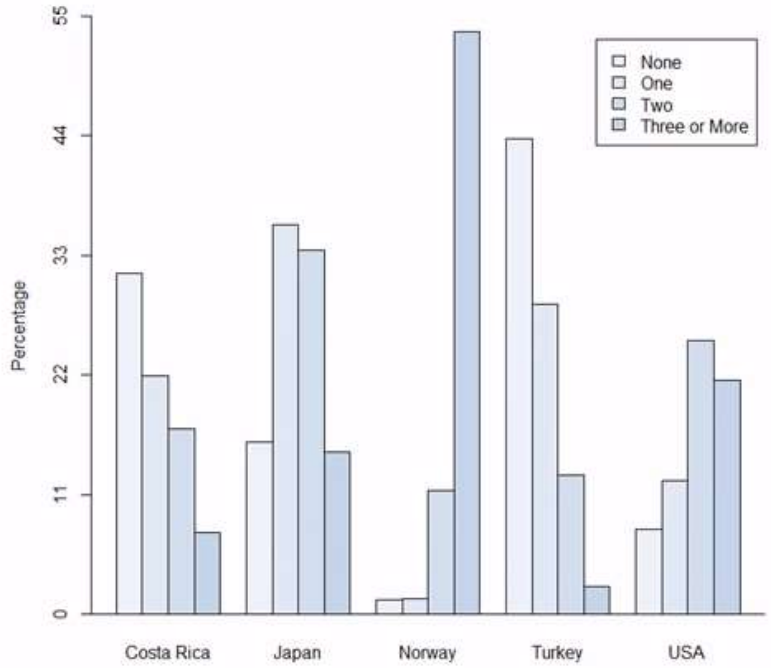
Figure 11-2 Common types of relationship between two variables

- Two categorical variables

Table 5

Number of computers in home by country

Country	None	One	Two	Three+
Costa Rica	31.36	21.94	17.10	7.50
Japan	15.80	35.79	33.49	14.89
Norway	1.31	1.49	11.37	53.52
Turkey	43.67	28.49	12.84	2.59
USA	7.86	12.29	25.20	21.50



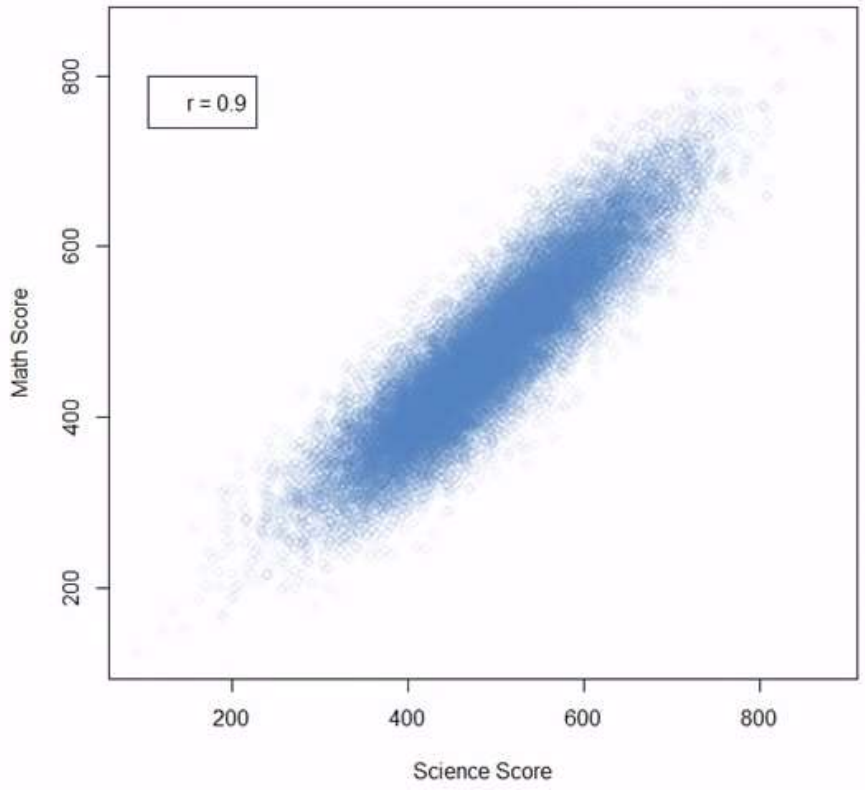
- The way we describe the relationship between two variables depends on the type of data each variable represents.
- Each participant must be measured on both variables.

- Two continuous variables

Table 6

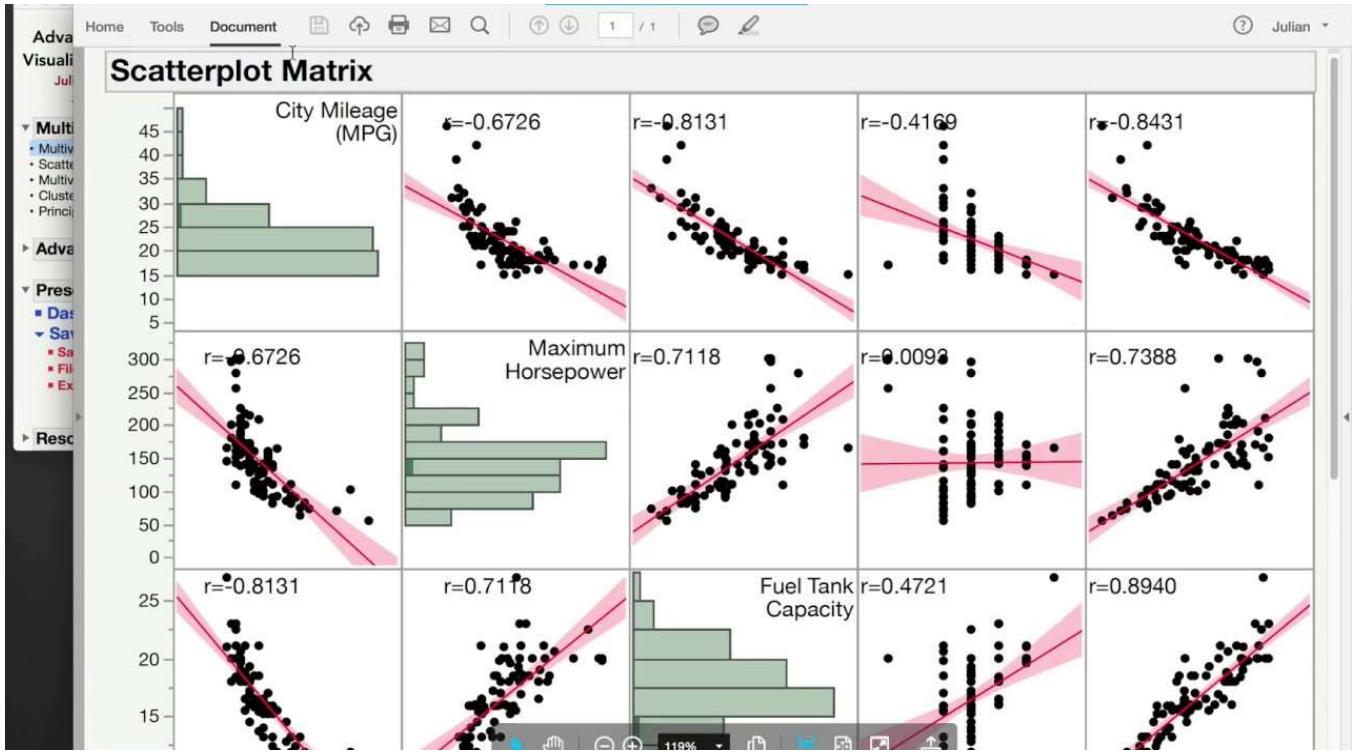
PISA test score

Score	Mean	S.D.
Science	491.47	97.03
Math	477.46	97.87



Example

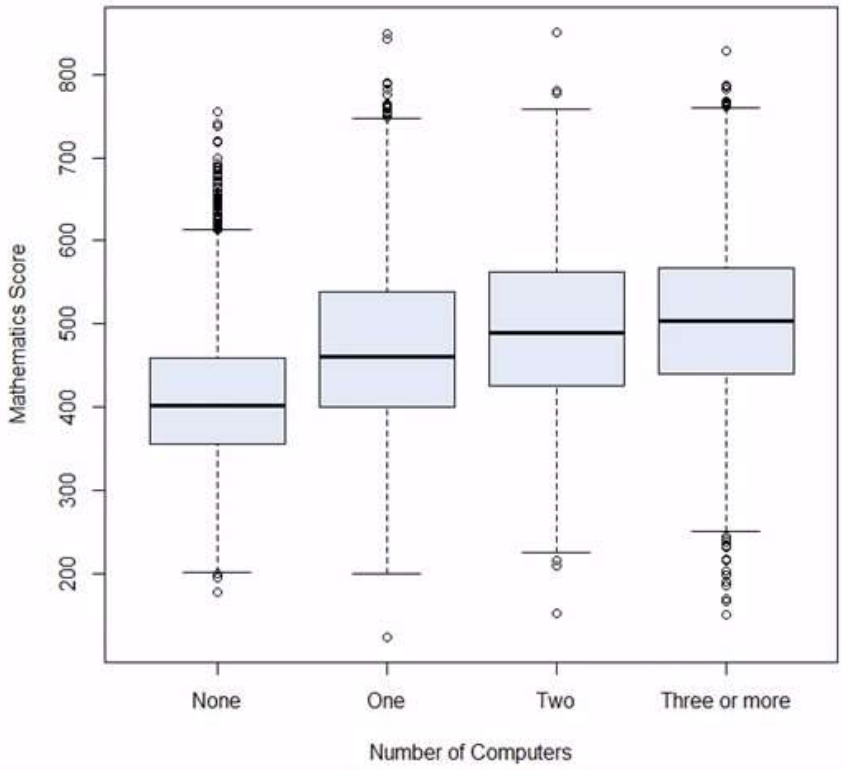
- Bi-variate Analysis



- One continuous and one categorical

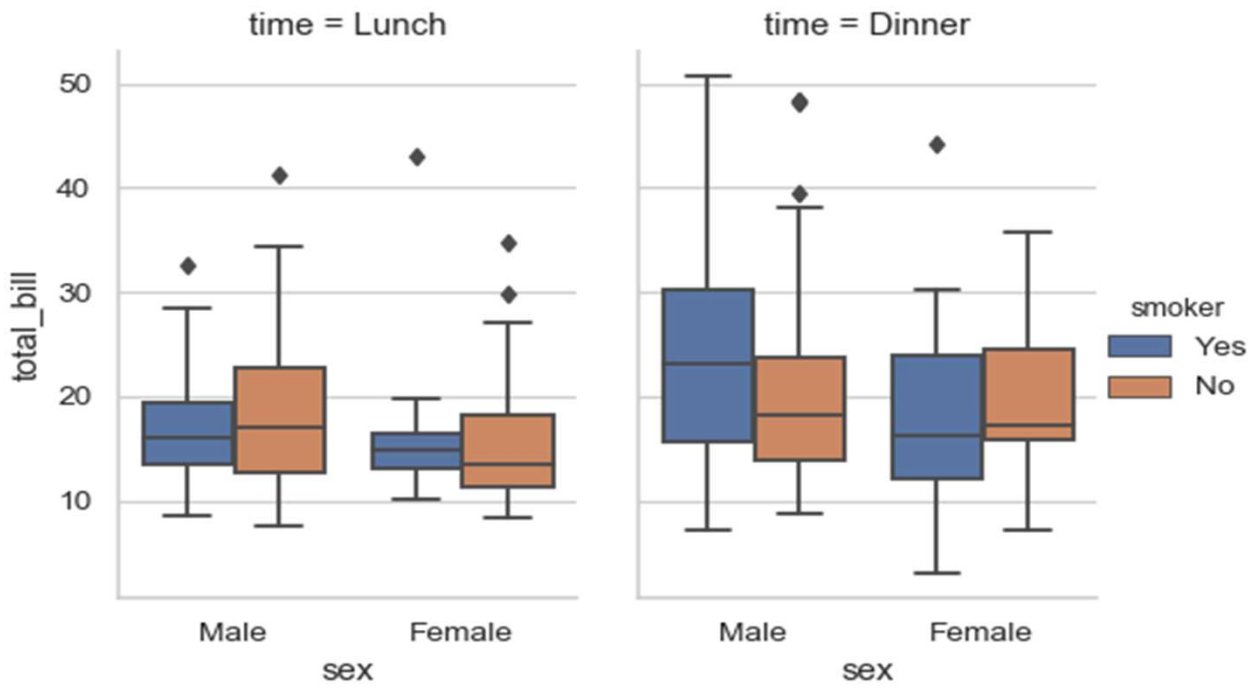
Table 7
Mathematics score by number of computers

Group	Mean	S.D.
None	413.15	83.47
One	471.46	96.87
Two	494.93	94.91
Three+	504.12	91.84



Example

- Multi-variate Analysis



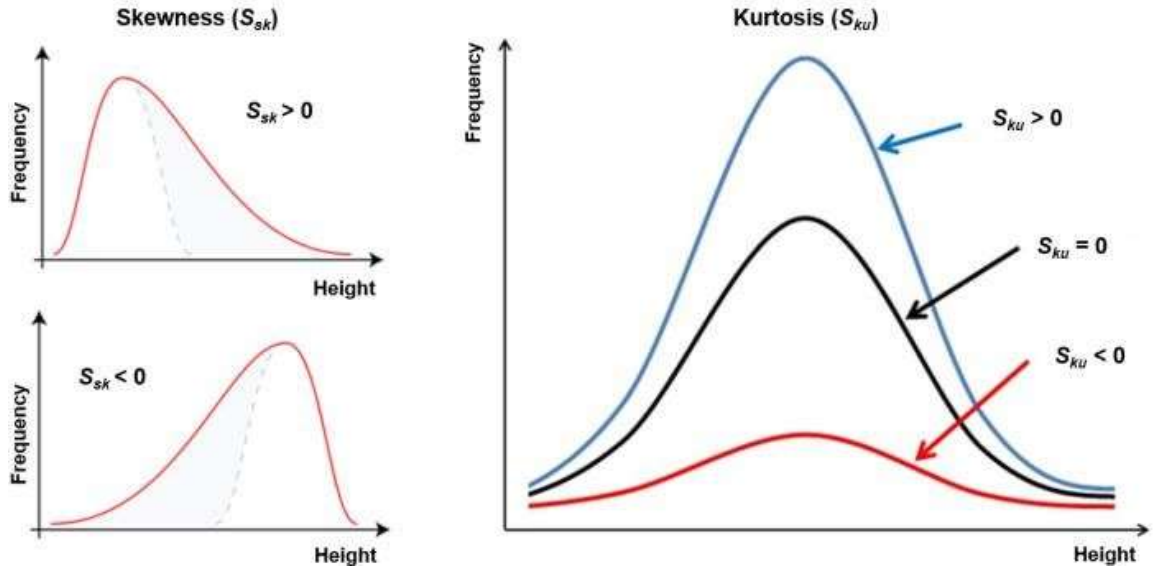
Skewness and Kurtosis

Moment number	Name	Measure of	Formula
1	Mean	Central tendency	$\bar{X} = \frac{\sum_{i=1}^{N} X_i}{N}$
2	Variance (Volatility)	Dispersion	$\sigma^2 = \frac{\sum_{i=1}^{N} (X_i - \bar{X})^2}{N}$
3	Skewness	Symmetry (Positive or Negative)	$Skew = \frac{1}{N} \sum_{i=1}^{N} \left[\frac{(X_i - \bar{X})}{\sigma} \right]^3$
4	Kurtosis	Shape (Tall or flat)	$Kurt = \frac{1}{N} \sum_{i=1}^{N} \left[\frac{(X_i - \bar{X})}{\sigma} \right]^4$

Where X is a random variable having N observations (i = 1,2,…,N).

Skewness and Kurtosis

Skewness, in basic terms, implies off-centre, so does in statistics, it means lack of symmetry. With the help of skewness, one can identify the shape of the distribution of data. **Kurtosis**, on the other hand, refers to the pointedness of a peak in the distribution curve. The main difference between skewness and kurtosis is that the former talks of the degree of symmetry, whereas the latter talks of the degree of peakedness, in the frequency distribution.



Outliers

- Outliers
  - Values smaller than lower inner fence, $Q_1 - 1.5IQR$
  - Values larger than upper inner fence, $Q_3 + 1.5IQR$
- Extreme values
  - Values smaller than lower outer fence, $Q_1 - 3IQR$
  - Values larger than upper outer fence, $Q_3 + 3IQR$
- Apply to continuous data

Effect of Outliers

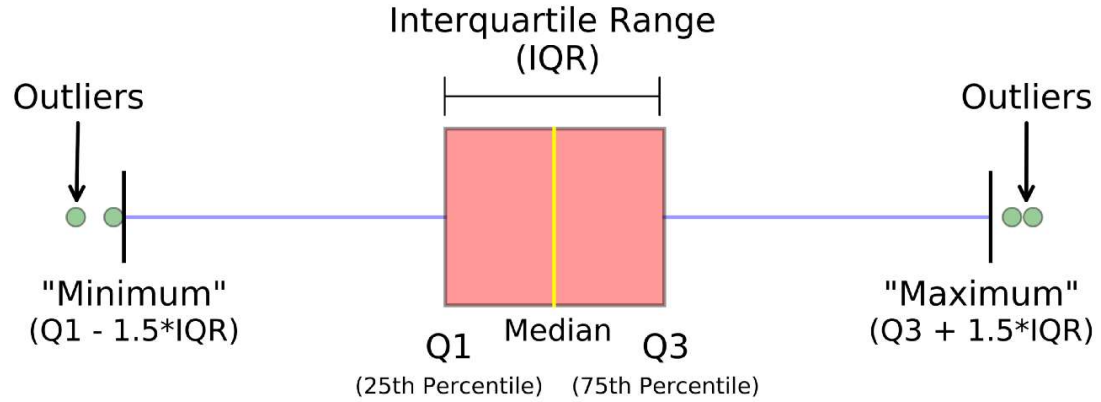
For the values 7,4,6,5,6,5,3,3,9,8

- Mean = 5.6
- Standard deviation = 2.01
- Median = 5.5
- IQR = 2.5

- For the values 7,4,6,5,6,5,3,3,20,8
  - Mean = 6.7 (was 5.6)
  - Standard deviation = 4.94 (was 2.01)
  - Median = 5.5
  - IQR = 2.5
- Now replace one value with an outlier

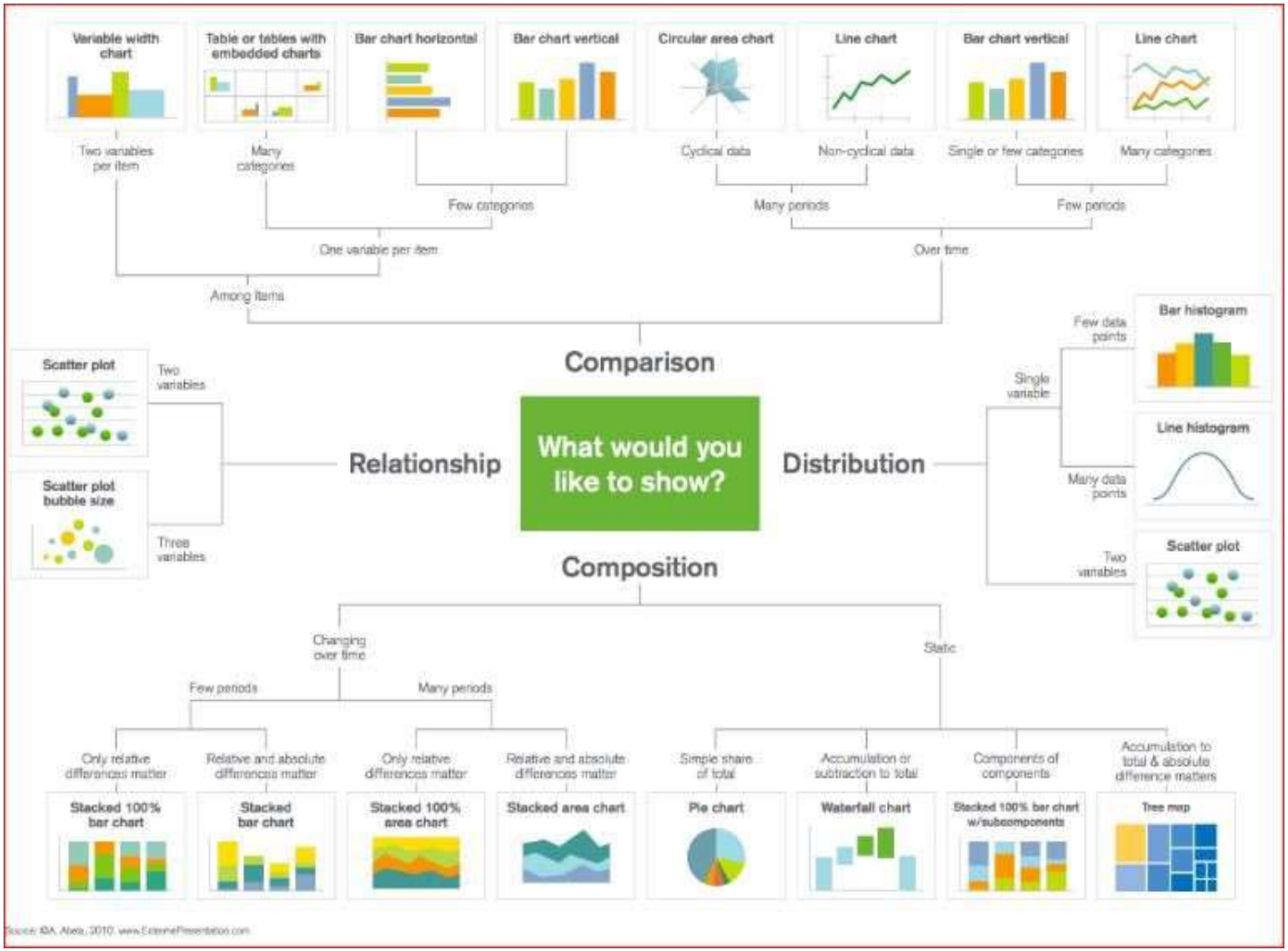
Robust Statistics for Continuous Data

- Median for central tendency
  - 50^{th} percentile
  - Second quartile, Q_2
- Interquartile range (IQR) for variability
  - IQR is the difference between the 75^{th} percentile and the 25^{th} percentile.
  - $IQR = Q_3 - Q_1$



- How do I know when to use the median instead of the mean?
- When should I use the IQR instead of the standard deviation?
- When should I use a boxplot instead of a bar chart?

Visualization

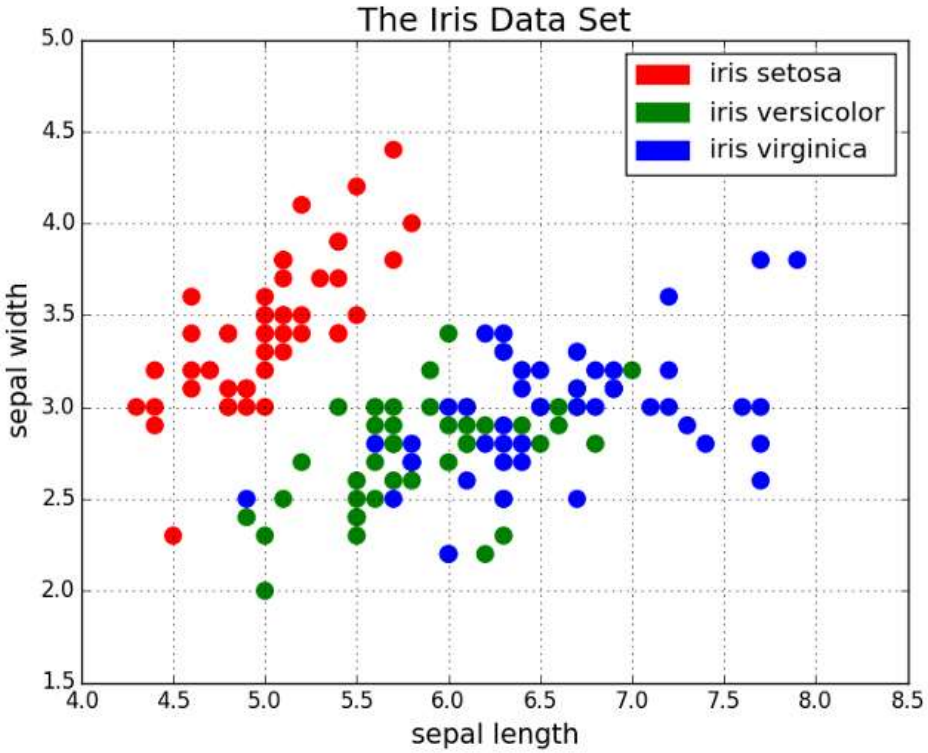


Visualization

- **Data Visualization** is a big part of a data scientist's jobs. In the early stages of a project, you'll often be doing an Exploratory Data Analysis (EDA) to gain some insights into your data.
- Creating visualizations really helps make things clearer and easier to understand, especially with larger, high dimensional datasets. Towards the end of your project, it's important to be able to present your final results in a clear, concise, and compelling manner that your audience, whom are often non-technical clients, can understand.

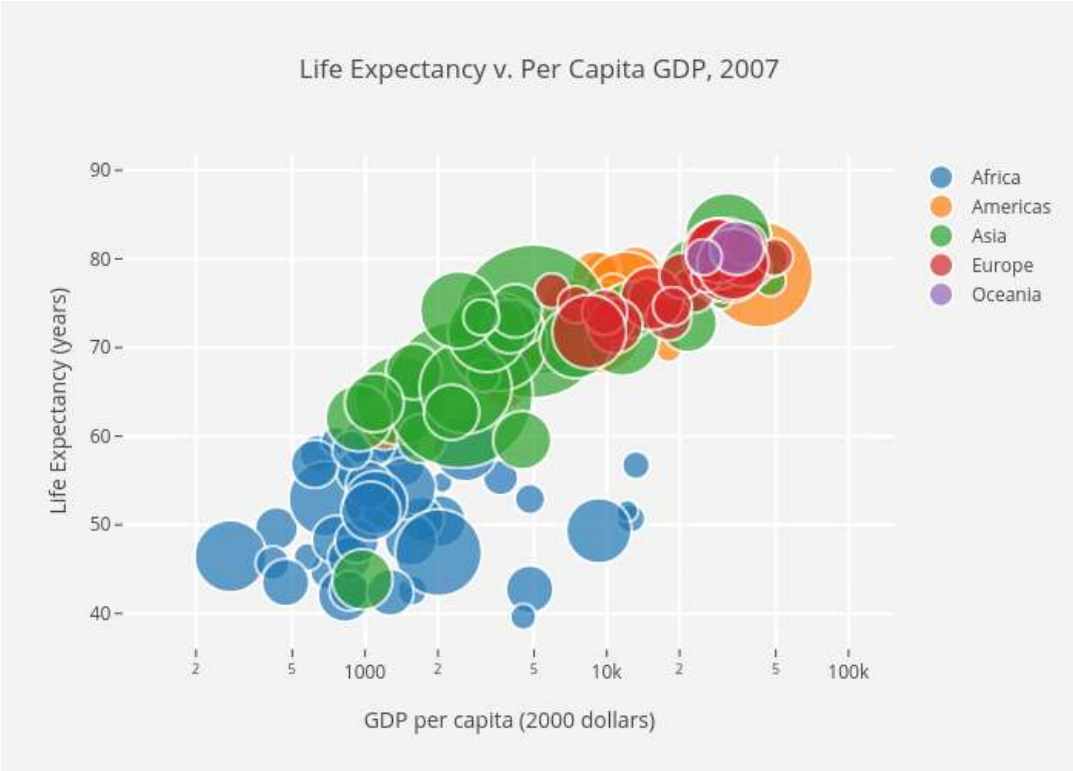
Visualization

- **Scatter Plots**
- Scatter plots are great for showing the relationship between two variables since you can directly see the raw distribution of the data. You can also view this relationship for different groups of data simple by colour coding the groups as seen in the first figure below.



Visualization

- **Scatter Plots**
- Want to visualize the relationship between three variables? No problemo! Just use another parameters, like point size, to encode that third variable as we can see in the second figure below. All of these points we just discussed also line right up with the first chart.

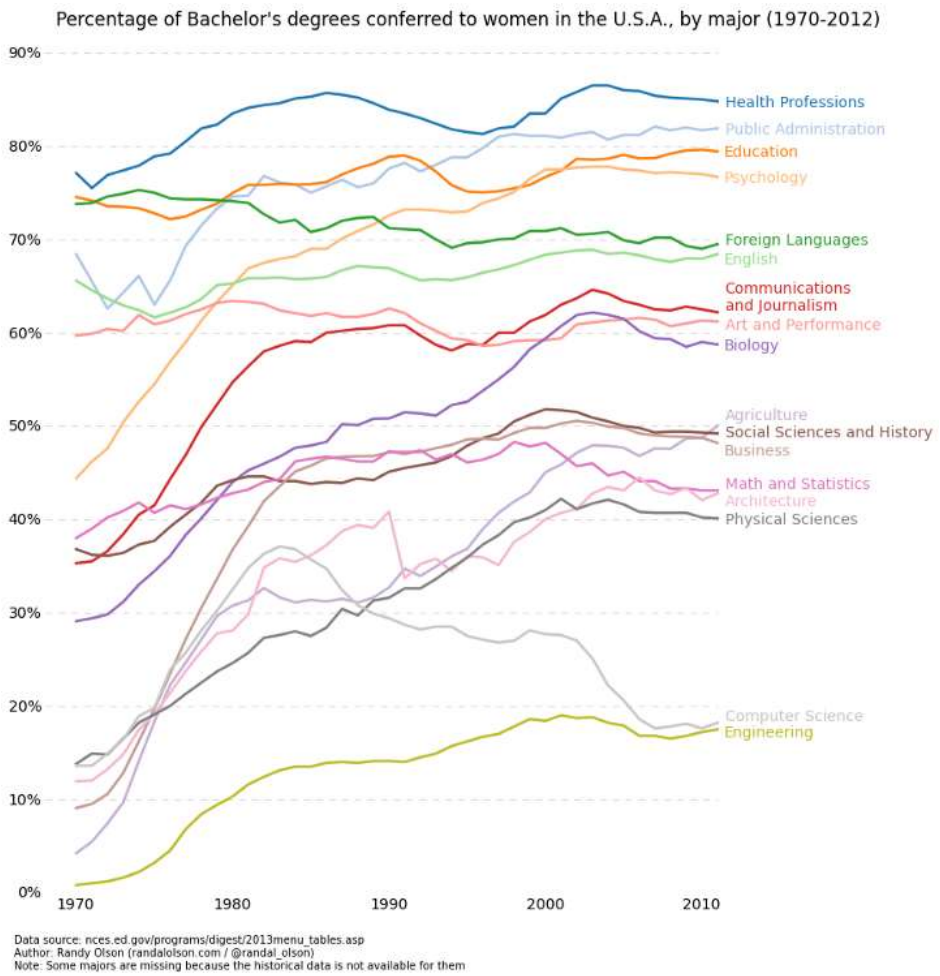


Visualization

- **Line Plots**
- Line plots are best used when you can clearly see that one variable varies greatly with another i.e they have a high covariance. Lets take a look at the figure below to illustrate. We can clearly see that there is a large amount of variation in the percentages over time for all majors.
- Plotting these with a scatter plot would be extremely cluttered and quite messy, making it hard to really understand and see what's going on. Line plots are perfect for this situation because they basically give us a quick summary of the covariance of the two variables (percentage and time).
- Again, we can also use grouping by colour encoding. Line charts fall into the "over-time" category from our first chart.

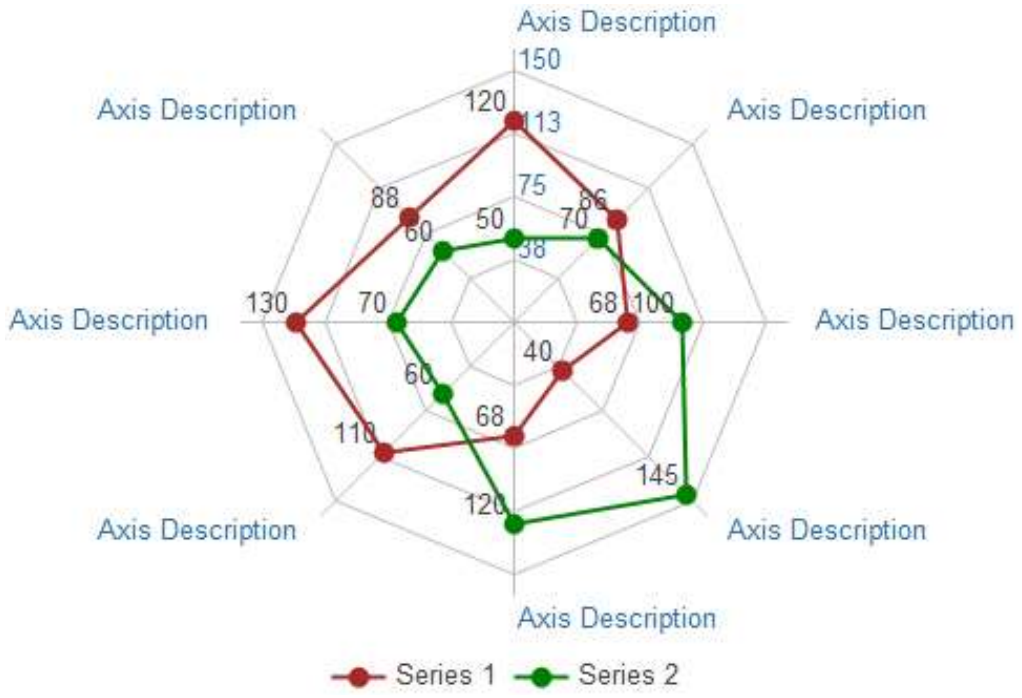
Visualization

- **Line Plots**



Visualization

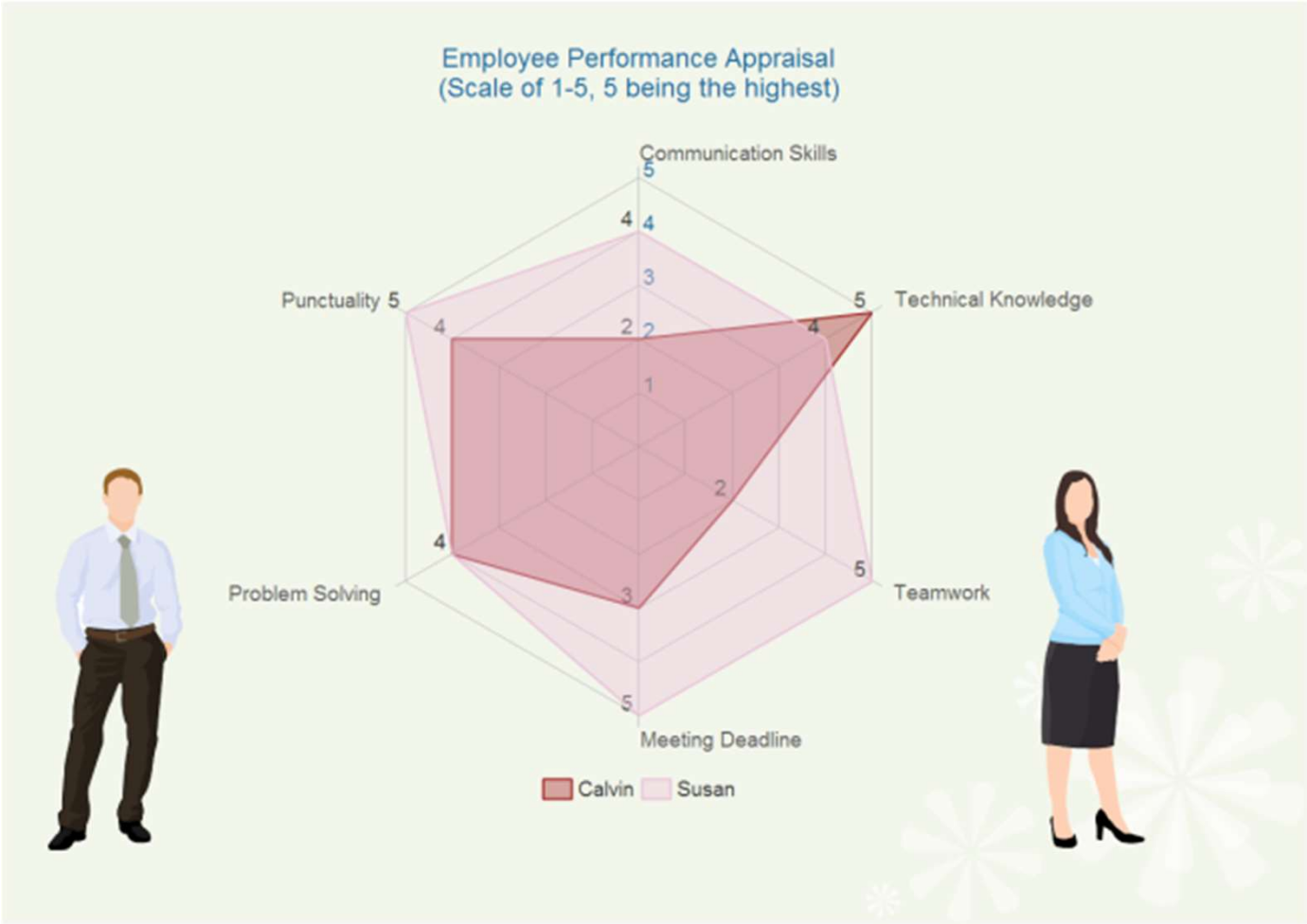
- **Spider and spider with markers**.
- With or without markers for individual data points, radar charts display changes in values relative to a center point.



Visualization

- **Filled spider**. In a filled spider chart, the area covered by a data series is filled with a color. Different series are filled with different colors.

Visualization



Visualization

Business Applications of Spider Chart
Employee Performance Appraisal

A spider chart can come in handy in the appraisal and review process of employees' performance. Here is an example appraising two employees' performance from the perspectives of Communication skills, punctuality, problem solving, meeting deadline, teamwork and technical knowledge.

As shown by the above spider chart, HR managers can visualize employee performance data, based on rankings offered by their respective seniors, on a single chart. This is very beneficial for HR management. This chart can also be applied to plan employee training by grouping employees who lack a particular skill set (low in rank) and then designing suitable remedial procedures for the group.

Visualization

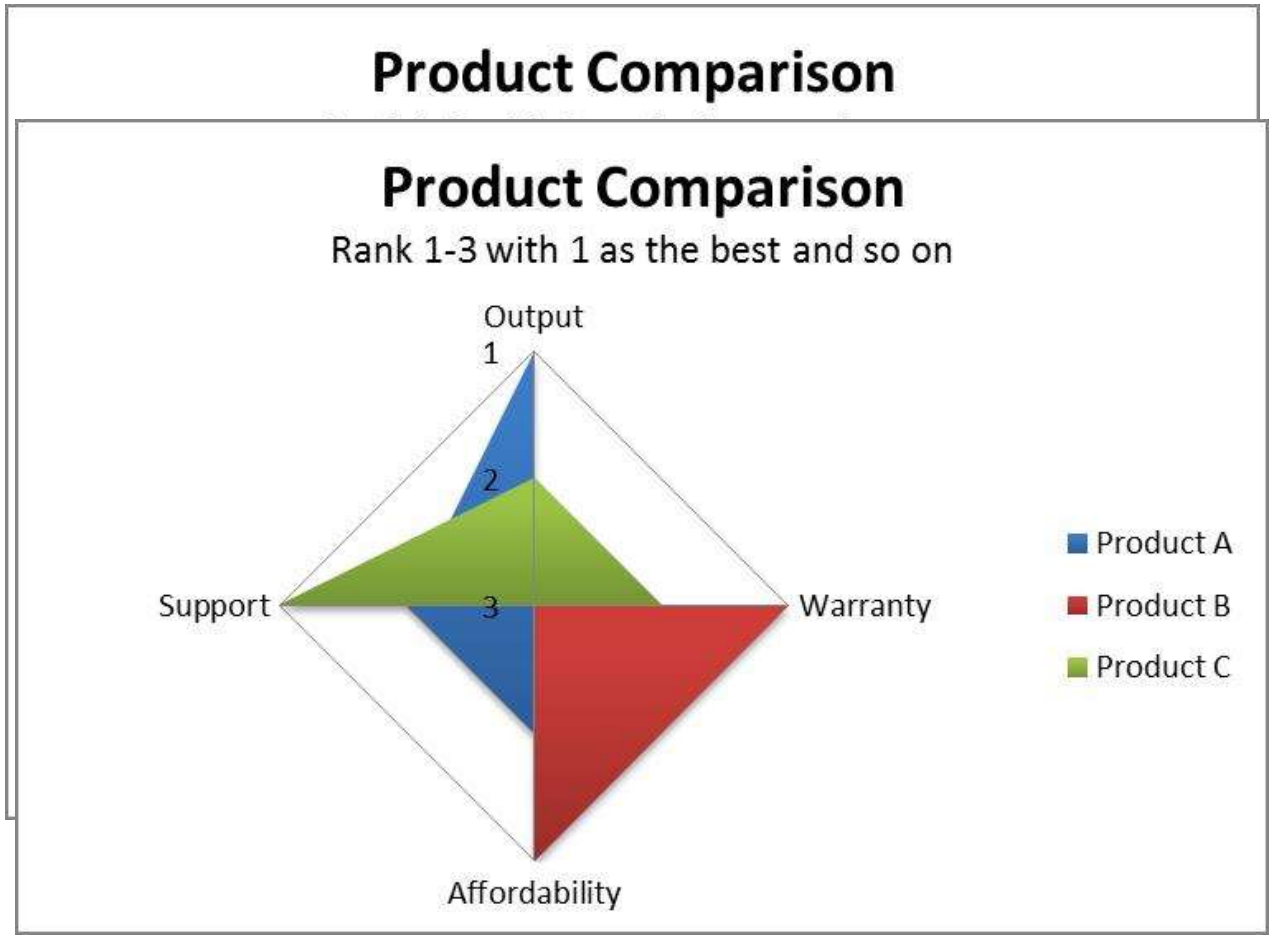
Product Comparison

Companies or consumers can use spider charts to compare products in terms of functions to highlight one product or choose the best product respectively. Let's say you want to compare 2 brands of smart phones over features like Battery, Camera, Display and Memory. Spider chart helps you get to know which brand is better when it comes to your most desired feature.

Sales Report

The sellers or salesmen can draw a spider chart to compare sales data of different goods. In this way, the most popular kind of goods can be figured out.

Visualization



Dashboard


1.198.198
VISITS
68 sec
AVG. VISIT DURATION
1.302.941
PAGE VIEWS
58,53 %
BOUNCE RATE
1,12
PAGES PER VISIT
Sessions by Week of Year
100k
50k
0k
Traffic Sources
17 % Direct
0 % Display
51 % Organic Search
9 % Paid Search
5 % Referral
19 % Social
Conversion Rates
Channel (Top 3 vs. Bottom 3)
Referral 23 %
Organic 23 %
Paid Se... 20 %
Campaign (Top 3 vs. Bottom 3)
Campai... 23 %
Campai... 22 %
Campai... 19 %
Landing Page (Top 3 vs. Bottom 3)
Page 2 26 %
Page 4 25 %
Page 3 22 %
Visitors by User Type
69 %
31 %
New
Returning
Bounce Rate by Week of Year
70 %
60 %
50 %
40 %

THANK YOU

Dr. Tran Anh Tuan, Faculty of Mathematics and Computer Science, University of Science, HCMC